FusionCatcher – a tool for finding somatic fusion genes in paired-end RNA-sequencing data

Daniel Nicorici*[1], Mihaela Şatalan[2], Henrik Edgren[3], Sara Kangaspeska[3], Astrid Murumägi[3], Olli Kallioniemi[3], Sami Virtanen[1], and Olavi Kilkku[1]

[1]Orion Corporation, P.O. Box 65, FI-02101 Espoo, Finland
[2]Florida Hospital Heartland Medical Center, P.O. Box 9400, Sebring, Florida 33871-9400, USA
[3]Institute for Molecular Medicine Finland, P.O. Box 20, FI-00014 Helsinki, Finland

November 19, 2014

Abstract

FusionCatcher is a software tool for finding somatic fusion genes in paired-end RNA-sequencing data from human or other vertebrates. FusionCatcher achieves competitive detection rates and real-time PCR validation rates in RNA-sequencing data from tumor cells. FusionCatcher is available at http://code.google.com/p/fusioncatcher/.

1 Introduction

Fusion genes are regarded as one of the common driver events behind tumor initiation and progression. Several fusion genes have also been found to be recurrent in multiple cancers. For example DNAJB1-PRKACA fusion was found, using FusionCatcher, in 100% of fibrolamellar hepatocellular carcinoma

*to whom correspondence should be addressed, email: Daniel.Nicorici@gmail.com

patients [15]. Also, fusion genes are good targets for therapeutic applications and personalized medicine[14].

One of the challenges in finding fusion genes in RNA-sequencing dataset is the very high rate of false positives which makes challenging their validation in the wet-lab. For example ChimeraScan[20] finds 28,399 fusion genes in four cancer breast cell lines[18] whilst:

- Cancer Genome Project lists 7,150 known fusion genes[1][22, 21], and
- Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer lists 2,094 known fusion genes[2][24].

in thousands of cancer cells/patients.

Throughout the paper we refer to somatic fusion genes, or shorter fusion genes, as the fusion genes which are found in majority of times in diseased cells/samples/patients. Therefore the expectation here is that number of fusion genes found in healthy samples is zero or very close to zero.

Thus the main goals of FusionCatcher are: (i) good real-time PCR validation rate (i.e. precision) which makes practical the validation of candidate fusion genes, and (ii) good detection rate (i.e. sensitivity) of fusion genes.

Here, we present FusionCatcher, which is a software tool for finding novel and known somatic fusion genes in paired-end RNA-sequencing data in diseased samples from vertebrates which have annotation data available in Ensembl database[10].

2 Methods

First, on the RNA-sequencing data is performed some pre-processing and filtering. Quality filtering of reads is performed, by:

- removing reads which align on ribosomal/transfer RNA, mitochondrial DNA, HLA genes, or known viruses/phages/bacteria genomes,
- trimming the reads which contain adapters and poly-A/C/G/T tails,
- clipping the reads based on quality scores, and
- removing the reads which are marked as bad quality by Illumina sequencer.

[1] November 2014
[2] August 2014

In an RNA-sequencing experiment, the RNA (which is converted to complementary DNA) is sequenced usually using next-generation sequencing platforms. Therefore, FusionCatcher is performing most of the data analysis at the RNA level (i.e. transcriptome) by aligning the sequencing reads, as single reads, on transcriptome using Ensembl genome annotation[10] and Bowtie aligner[8]. Furthermore, the reads are mapped on genome, using the Bowtie aligner, for filtering purposes and the reads which have a better alignment (i.e. fewer mismatches) at the genome level will have their transcriptome mappings removed. Similarly, the reads, which map simultaneously on several transcripts of different genes, have their mappings on transcriptome removed, and for the corresponding genes a sequence similarity score is computed using these read counts. The unmapped reads, which are the reads which passed the quality filtering and do not map on the transcriptome and the genome, are kept for further analyses.

The reads mapping on the transcriptome are used further to build a preliminary list of candidate fusion genes by searching for pairs of genes, such that for each pair of genes (gene A, gene B) one read maps on gene A's transcripts and its paired-read maps on gene B's transcripts. From the preliminary list of candidate fusion genes are removed the pairs of genes, using known and novel criteria, which make biological sense, such as:

- both genes are known to be the other's paralog in Ensembl database,
- a gene is known to be the other's pseudogene in Ensembl database,
- a gene is known micro/transfer/Y/7SK/small-nuclear/small-nucleolar RNA,
- fusion is known *a priori* to be a false positive event (e.g. ConjoinG database [12], conjoined HLA genes [13],
- it has been found previously in samples from healthy persons, like for example from Illumina Body Map 2.0 RNA-sequencing data[23] and an in-house RNA-sequencing database of healthy samples,
- both genes are overlapping each other on the same strand according to one of the public known databases, such as Ensembl, UCSC, or RefSeq databases, and
- pair of genes which have a very high count of reads mapping simultaneously on both of genes forming a fusion.

FusionCatcher uses and ensemble approach consisting of four different methods and four different aligners for identifying the fusion junctions. Each

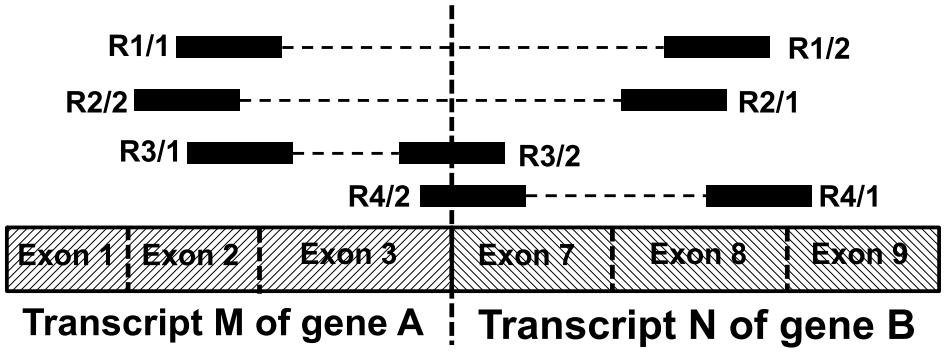

Figure 1: Bowtie's mappings of reads and their corresponding paired-reads which support the fusion between genes A and B with fusion junction between exon 3 of gene A and exon 7 of gene B. The pairs of reads (R1/1, R1/2) and (R2/1, R2/2) support the fusion and the reads R3/2 and R4/2 support the exon-exon junction of the fusion.

method corresponds to one aligner and the aligners are Bowtie[8], BLAT[9], STAR[16], and Bowtie2[17].

For the first method, which is using information regarding the exon/intron positions (i.e. genome annotation), a database of exon-exon junctions is built, and it contains all the exon-exon junctions for all possible exon-exon combinations for each candidate fusion gene. The unmapped reads are aligned on the database using Bowtie aligner. An unmapped read is counted as evidence for supporting a candidate fusion gene if it is found to (i) map on a exon-exon junction belonging to a candidate fusion gene (i.e. reads R3/2 and R4/2 in Figure 1), and (ii) have its corresponding paired-read mapping on one of the genes forming this candidate fusion gene (i.e. reads R3/1 and R4/1 in Figure 1). The unmapped reads which have the same mapping position on the same exon-exon junction are counted only once. Therefore, the candidate fusion genes which have been found to have counts of such unmapped reads larger than a given threshold will make it to the final list of fusion genes. The role of the first method is to reduce the number of unmapped reads given as input to the next three methods which are more computationally demanding. Also the unmapped reads found to map here are removed from further analysis.

For the next three methods, which are not making use of exon/intron positions (only information regarding start and end positions of genes is used), another database of gene-gene sequences is built (e.g. gene A – gene B shown in Figure 2). The database contains all the gene-gene sequences for each candidate fusion gene. The unmapped reads, which still remain unmapped after aligning them on the exon-exon junctions database using the

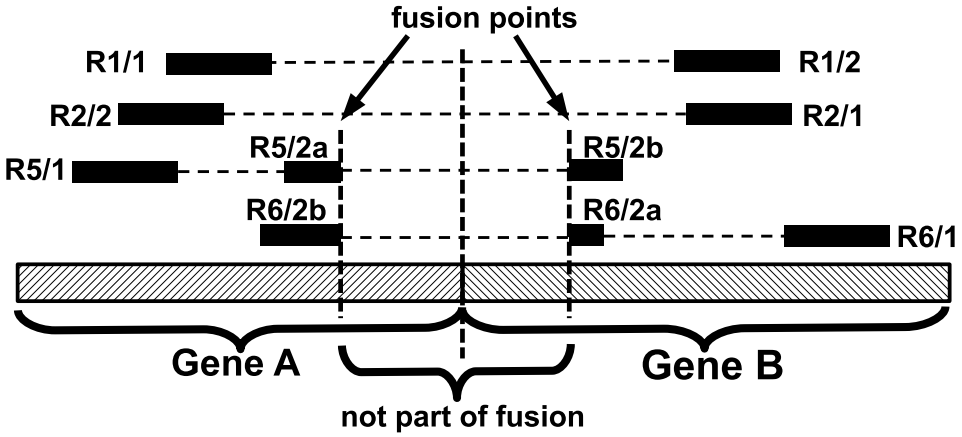

Figure 2: Mappings of reads and their corresponding paired-reads which support the fusion between genes A and B. The pairs of reads (R1/1, R1/2) and (R2/1, R2/2) support the fusion and the reads R5/2 and R6/2 support the junction of the fusion. BLAT, STAR, and Bowtie2 aligners split the read R5/2 into two parts R5/2a and R5/2b and the read R6/2 into R6/2a and R6/2b. The reads R1/1, R1/2, R2/1, R2/2, R5/1, and R6/1 are mapped using Bowtie aligner on transcriptome and the reads R5/2 and R6/2 are mapped using BLAT, STAR, and Bowtie2 aligners on gene sequences.

first method, together with the reads, which are supporting the candidate fusion genes, are aligned further on this database using BLAT aligner[9], STAR aligner[16], and Bowtie2 aligner[17]. A read is counted as evidence for supporting the candidate fusion gene if it is found to (i) map on a gene-gene sequence belonging to a candidate fusion gene such that first part of the read maps on first gene and the second part of the read maps on the second gene (i.e. reads R3/2 and R4/2 in Figure 2), and (ii) have its corresponding paired-read mapping in the transcriptome on one of the genes forming this candidate fusion gene (i.e. reads R4/2 and R5/2 in Figure 2). The reads which have the same mapping positions on the same gene-gene sequence are counted only once, because different reads which share the same start and end position are most likely artefact of PCR process used during sample preparation and sequencing. Therefore, the candidate fusion genes which have been found to have counts of such reads over a given threshold will make it to the final list of fusion genes. The final list of candidate fusion genes encompasses the lists of candidate fusion genes found using all four methods.

FusionCatcher is written in Python, runs on Linux, and it is freely available under the GPL version 3 license at http://code.google.com/p/fusioncatcher/.

3 Results and Discussion

The tumor cell lines SNU-16, KATOIII, and NCI-H716 are known to harbor FGFR2 amplifications. As fusion genes are known to reside on re-arranged areas on the genome, this indicates that the FGFR2 gene is probably a fusion gene partner [11, 14] in these cell lines. This hypothesis is confirmed *in silico* by FusionCatcher which found several novel FGFR2 fusions:

- SNU-16 cell line: FGFR2-CD44, FGFR2-PPAPDC1A, FGFR2-MYC, and FGFR2-PDHX [3],
- KATOIII cell line: FGFR2-CEACAM5, FGFR2-ULK4, GCNT3-FGFR2, SNX19-FGFR2, CTNNB1-FGFR2[4], and
- NCI-H716 cell line: COL14A1-FGFR2, FGFR2-COL14A1, PVT1-FGFR2[5].

Additional novel fusion genes have been detected by FusionCatcher in HeLa and U87MG cell lines [6, 7].

	Validated fusion genes	Cells	FusionCatcher*	FusionCatcher**	SOAPFuse	ChimeraScan	deFuse	FusionHunter	SnowShoes-FTD	TopHat-Fusion
1	ACACA-STAC2†	BT-474	x	x	x	x	x	x	x	x
2	RPS6KB1-SNF8†	BT-474	x	x	x	x	x	x	x	x
3	VAPB-IKZF3†	BT-474	x	x	x	x	x	x	x	x
4	ZMYND8-CEP250†	BT-474	x	x	x	x	x	x	x	x
5	RAB22A-MYO9B†	BT-474	x	x	x	x	x	–	x	x
6	SKA2-MYO19†	BT-474	x	x	x	x	x	x	–	x
7	DIDO1-KIAA0406†	BT-474	x	–	x	x	x	–	–	x
8	STARD3-DOK5†	BT-474	x	x	x	x	x	x	x	x
9	LAMP1-MCF2L†	BT-474	x	–	x	x	x	–	x	–
10	GLB1-CMTM7†	BT-474	x	x	x	x	x	x	x	x
11	CPNE1-PI3†	BT-474	x	–	–	x	–	–	–	x
12	THRA-AC090627.1	BT-474	x	x	x	–	x	–	–	x
13	TOB1-SYNRG	BT-474	x	x	x	x	x	x	x	x
14	AHCTF1-NAAA	BT-474	x	x	x	–	x	–	–	x
15	MED1-STXBP4	BT-474	x	x	x	x	x	–	x	x
16	MED13-BCAS3	BT-474	x	x	x	x	x	x	x	x
17	MED1-ACSF2	BT-474	x	x	x	x	x	–	x	x
18	TRPC4AP-MRPL45	BT-474	x	x	x	–	x	x	x	x
19	STX16-RAE1	BT-474	x	x	x	x	x	x	–	x
20	USP32-MED1†	BT-474	x	x	x	x	–	–	–	–
21	PIP4K2B-RAD51C	BT-474	x	–	x	x	–	–	–	x
22	TATDN1-GSDMB†	SK-BR-3	x	x	x	x	x	x	x	x
23	CSE1L-AL035685.1†	SK-BR-3	x	–	x	–	x	–	–	–
24	RARA-PKIA†	SK-BR-3	x	x	x	x	x	x	x	x
25	ANKHD1-PCDH1†	SK-BR-3	x	x	x	x	x	x	x	x
26	CCDC85C-SETD3†	SK-BR-3	x	x	x	x	–	–	x	–
27	SUMF1-LRRFIP2†	SK-BR-3	x	x	x	x	x	–	x	x
28	WDR67-ZNF704†	SK-BR-3	x	x	x	x	–	–	x	x
29	CYTH1-EIF3H †	SK-BR-3	x	x	x	x	x	x	–	x
30	DHX35-ITCH†	SK-BR-3	x	–	x	x	–	–	–	–
31	NFS1-PREX1†	SK-BR-3	x	–	x	–	x	–	–	–
32	BSG-NFIX†	KPL-4	x	x	x	x	x	x	x	x
33	PPP1R12A-SEPT10†	KPL-4	x	x	x	x	x	–	x	x
34	NOTCH1-NUP214†	KPL-4	x	x	x	x	x	x	x	x
35	BCAS4-BCAS3	MCF-7	x	x	x	x	x	x	x	x
36	ARFGEF2-SULF2	MCF-7	x	x	x	x	x	x	x	x
37	RPS6KB1-TMEM49	MCF-7	x	x	x	x	x	–	–	–
38	GCN1L1-MSI1	MCF-7	x	–	x	x	–	–	–	–
39	AC099850.1-TMEM49	MCF-7	x	x	x	–	x	–	–	x
40	SMARCA4-CARM1	MCF-7	x	x	x	x	x	–	x	–
	COUNTS of validated fusions		**40**	32	**39**	**34**	33	19	25	31
	COUNTS of reported fusions		43	38	65	28399	163	39	33	108
	Precision		**0.93**	**0.84**	0.60	0.001	0.20	0.49	0.76	0.29
	False Discovery Rate		**0.07**	**0.16**	0.40	0.99	0.79	0.51	0.24	0.71

Table 1: Comparisons of RNA-seq fusion finders on breast cancer cells dataset[2]. FusionCatcher* is version 0.91 (i.e. Ensembl release 61 and GRCh37/hg19) from[1, 2]. FusionCatcher** is version 0.99.3e (i.e. Ensembl release 77 and GRCh38/hg38). Data regarding SOAPFuse, ChimeraScan, deFuse, FusionHunter, SnowShoes-FTD, and TopHat-Fusion (GRCh37/hg19) are from[18]. Fusion genes marked with † have been found for the first time by FusionCatcher*[1, 2]

	Spike-in fusion gene	Replicate	FusionCatcher**	TopHat-Fusion	ChimeraScan	SnowShoes-FTD
1	EWSR1-ATF1	R1	x	x	x	x
2	EWSR1-ATF1	R2	x	–	x	x
3	TMPRSS2-ETV1	R1	x	–	x	–
4	TMPRSS2-ETV1	R2	x	–	x	–
5	EWSR1-FLI1	R1	x	x	x	x
6	EWSR1-FLI1	R2	x	–	x	x
7	NTRK3-ETV6	R1	x	x	x	x
8	NTRK3-ETV6	R2	x	x	x	x
9	CD74-ROS1	R1	x	x	x	x
10	CD74-ROS1	R2	x	x	x	x
11	HOOK3-RET	R1	x	x	x	x
12	HOOK3-RET	R2	x	x	x	x
13	EML4-ALK	R1	x	–	x	–
14	EML4-ALK	R2	x	x	x	–
15	AKAP9-BRAF	R1	x	x	x	x
16	AKAP9-BRAF	R2	x	x	x	x
17	BRD4-NUT	R1	x	x	x	x
18	BRD4-NUT	R2	x	–	x	x
	COUNTS of spike-in fusions		**18**	12	**18**	14

Table 2: Comparisons of fusion genes finders on synthetic spike-in fusion genes dataset (concentration [log10 (pMol)] = -8,57; minimum fusion-supporting read cut off = 2)[19]. FusionCatcher** is version 0.99.3e (i.e. Ensembl release 77 and GRCh38/hg38). Data regarding TopHat-Fusion, ChimeraScan, and SnowShoes-FTD (GRCh37/hg19) are from[19].

The breast cancer RNA-sequencing dataset from [2] contains 40 real-time PCR validated fusion genes, of which 27 are published in [2] and 13 in [1]. All 40 fusion genes were detected by FusionCatcher and 25 (marked with †in Table 1) were found for the first time by FusionCatcher in[1, 2]. In Table 1 is presented a comparison of several fusion genes finders like SOAPFuse[18], ChimeraScan[20], deFuse[27], FusionHunter[28], SnowShoes-FTD[26], and TopHat-Fusion[26] using the same dataset[18]. Also FusionCatcher is run (see Table 1) on the same data set using two different Ensembl genome annotations, which are release 61 (based on GRCh38/hg38 assembly) and release 77 (based on GRCh3/hg19 assembly). As expected the genome annotations and genome assemblies used, affect the peformance of fusion genes finding. As shown in Table 1, FusionCatcher has the best precision, i.e. $TP/(TP+FP)$, and also competitive sensitivity irrespective to the genome annotation and assembly used.

In Table 2[3] is shown another comparison of fusion genes finders on a synthetic spike-in fusion genes dataset from [19], where FusionCatcher detects all the synthetic fusion genes.

In summary, FusionCatcher has the best precision, which translates into in real-time PCR validation rates, for finding fusion genes in the breast cancer RNA-sequencing data set[1, 2, 18] and very good sensitivity in synthetic spike-in fusion genes dataset[19].

References

[1] S. Kangaspeska, et al, *Reanalysis of RNA-sequencing data reveals several additional fusion genes with multiple isoforms*, PLoS ONE, **7(10)**, 2012.

[2] H. Edgren, et al, Identification of fusion genes in breast cancer by paired-end RNA-sequencing, Genome Biology, **12:R6**, 2011.

[3] D. Nicorici, *Novel FGFR2 fusion genes in SNU-16 gastric cancer cell line*, Figshare, 2013, DOI:10.6084/m9.figshare.856657

[4] D. Nicorici, *Novel FGFR2 fusion genes in KATOIII gastric cancer cell line*, Figshare, 2013, DOI:10.6084/m9.figshare.856658

[5] D. Nicorici, *Novel FGFR2 fusion genes in NCI-H716 colorectal cancer cell line*, Figshare, 2014, DOI:10.6084/m9.figshare.1125933

[3]only samples having concentration of -8,57 are used here because all the other samples are missing the replicates data (November 2014)

[6] D. Nicorici, *Fusion genes in HeLa cervical cancer cell line*, Figshare, 2013, DOI:10.6084/m9.figshare.856664

[7] D. Nicorici, *Novel fusion genes in U87MG glioblastoma-astrocytoma cell line*, Figshare, 2013 DOI:10.6084/m9.figshare.856659

[8] B. Langmead, C. Trapnell, M. Pop, S.L. Salzberg, *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome*, Genome Biology, 10:R25, 2009.

[9] W.J. Kent, *BLAT – The BLAST-Like Alignment Tool*, Genome Research, 4, pp. 656–664, 2002.

[10] P. Flicek et al. *Ensembl 2013*, Nucleic Acids Research, 41, 2013, DOI:10.1093/nar/gks1236

[11] Y.M. Wu et al., Identification of Targetable FGFR Gene Fusions in Diverse Cancers, Cancer Discovery, 3, pp. 636–647, 2013, DOI:10.1158/2159-8290.CD-13-0050.

[12] T. Prakash et al. Expression of Conjoined Genes: Another Mechanism for Gene Regulation in Eukaryotes, PLoS ONE, 5(10), 2010, DOI:10.1371/journal.pone.0013284

[13] S. Nacu et al. *Deep RNA sequencing analysis of readthrough gene fusions in human prostate adenocarcinoma and reference samples*, BMC Medical Genomics, 4:11, 2011, DOI:10.1186/1755-8794-4-11

[14] X. Su et al. *FGFR2 amplification has prognostic significance in gastric cancer: results from a large international multicentre study*, British Journal of Cancer, 110, pp. 967–975, 2014, DOI:10.1038/bjc.2013.802

[15] J.N. Honeyman, et al. *Detection of a Recurrent DNAJB1-PRKACA chimeric transcript in fibrolamellar hepatocellular carcinoma*, Science, 343, pp. 1010–1014, 2014, DOI:10.1126/science.1249484

[16] A. Dobin et al., *STAR: ultrafast universal RNA-seq aligner*, Bioinformatics, 29, pp. 15–21, 2013, DOI:10.1093/bioinformatics/bts635

[17] B. Langmead, S. Salzberg, *Fast gapped-read alignment with Bowtie 2*, Nature Methods, 9, pp. 357–359, 2012, DOI:10.1038/nmeth.1923

[18] W. Jia et al., *SOAPfuse: an algorithm for identifying fusion transcripts from paired-end RNA-Seq data*, Genome Biology, 14:R12, pp. 2971–2978, 2013, DOI:10.1186/gb-2013-14-2-r12

[19] W.D. Tembe et al., *Open-access synthetic spike-in mRNA-seq data for cancer gene fusions*, BMC Genomics, 15:824, 2014, DOI:10.1186/1471-2164-15-824

[20] M.K. Iyer et al. *ChimeraScan: a tool for identifying chimeric transcription in sequencing data*, Bioinformatics, 27(20), pp. 2903-2904, 2011. DOI:10.1093/bioinformatics/btr467

[21] P.A. Futreal et al. *A census of human cancer genes*, Nature Reviews Cancer, 4, pp. 177-183, 2014. DOI:10.1038/nrc1299

[22] – *Cancer Genome Project*, http://www.sanger.ac.uk/research/projects/cancergenome/

[23] – *RNA-Seq of human individual tissues and mixture of 16 tissues (Illumina Body Map 2.0)*, http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-513/

[24] – *Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer*, Mitelman F, Johansson B and Mertens F (Eds.), 2014, http://cgap.nci.nih.gov/Chromosomes/Mitelman

[25] D. Kim D et al., *TopHat-fusion: an algorithm for discovery of novel fusion transcripts*, Genome Biology, 12:R72, 2011, DOI:10.1186/gb-2011-12-8-r72

[26] Y.W. Asmann et al., *A novel bioinformatics pipeline for identification and characterization of fusion transcripts in breast cancer and normal cell lines*, Nucleic Acids Research, 39:e100, 2011 DOI:10.1093/nar/gkr362

[27] A. McPherson et al., *deFuse: An algorithm for gene fusion discovery in tumor RNA-Seq data*, PLoS Computational Biology, 2011 DOI:10.1371/journal.pcbi.1001138

[28] Y. Li et al., *FusionHunter: identifying fusion transcripts in cancer using paired-end RNA-seq*, Bioinformatics, 27(12), pp. 1708–1710, 2011, DOI:10.1093/bioinformatics/btr265